Homework 1

Mark Schulist

1.

a. Each row represents an observational unit in the data set, in this case, they represent UK residents.

b. There were 1691 participants in the study—it's the number of rows in the data set.

c. sex: categorical, nominal; age: numerical, continuous; marital: categorical, nominal; grossIncome: categorical, ordinal; smoke: categorical, nominal; amtWeekends: categorical, ordinal; amtWeekdays: categorical, ordinal

2.

a. No, this is an observational study and not an experiment. They did not prescribe treatments to different randomized groups.

b. Maybe drinking more coffee makes people have more cramps. Less sleep could also make people get cramps. These are confounding variables that might indirectly be the cause of the relationship between stress and muscle cramps.

3.

a. You will get a nice sample from the area, but you might miss out on some areas more than others because of randomness. It could also be more expensive to go to 200 random households across the entire city than just a few neighborhoods. It takes a lot of effort to go to many spread out households.

b. Depending on how you split the neighborhoods, you might over represent certain neighborhoods. Certain neighborhoods might have larger populations, so you'd want to sample more from that area. Of course, this depends on the type of survey being conducted and if different neighborhoods are significantly different in their opinions on the survey. This will still be challenging to sample as you are going to many random homes across all neighborhoods.

c. You will miss out on $\frac{17}{20}$ of the neighborhoods, but you will get a good sample from the ones you sample. This strategy could be good if the neighborhoods you sample are representative of the rest of the city. It will also be easy to sample as you only visit 3 out of the 20 neighborhoods.

4. In the field of data science, there are many times that bias might occur. Let's say that, theoretically, I work for a company that collects data on the political viewpoints of users of a certain website, let's say a website that makes comics about computers and math. If we tried to use these samples to determine the outcome of an election, we would most likely get the wrong result, even if our sample size was large. The people using this website are most likely very academic and are probably more likely to be liberal than conservative. Trying to extrapolate the results of this sample to the population of Americans would not work very well due to the bias of the type of people who are interested in computers and math.

5.

a. (1) 3, 5, 6, 7, 9 (2) 3, 5, 6, 7, 20

They have the same median and IQR because `20` is the last number in the ordered list, and the median and IQR are not affected by the last and first numbers in a list. (2) will have a longer right tail (right skewed) because of the `20`, in fact it is considered an outlier.

b. (1) 3, 5, 6, 7, 9 (2) 3, 5, 7, 8, 9

(1) will have a slightly smaller IQR and median due to it having 2 smaller numbers. (1) will be normal whereas (2) will be slightly left skewed as there are more large numbers than small numbers.

c. (1) 1, 2, 3, 4, 5 (2) 6, 7, 8, 9, 10

They have the same IQR, but (2) has a higher median due to there being higher numbers. The shape will be the same because they are both 5 numbers with increasing by one.

d. (1) 0, 10, 50, 60, 100 (2) 0, 100, 500, 600, 1000

The median and IQR will be larger for (2) because the numbers are bigger and spread out more. Both will have the same shape because they are just scaled versions of each other.

6.

a. Normal, low of 50 and high of 70. `(2)`

b. Uniform, low of 0 and high of 100. `(3)`

c. Right skewed, low of 0 and high of 8. `(1)`

7.

a. Right skewed. Median as there is a lot of skew. IQR would be best as the standard deviation will be affected a lot by the super expensive homes. There are a lot of cheap homes, and few expensive ones.

b. Symmetric. Mean would be best as it is not skewed. Standard deviation would also be best as it is not skewed and symmetric. There are a similar number of expensive homes as cheap homes, so there is less skew than the previous question.

c. Left skewed. Most people drink few or zero, but a small number of people drink a LOT. Median and IQR because of the skew.

8.

```
survey <- matrix(c(57, 121, 179, 15, 120, 113, 126, 4, 101, 28, 45, 1),
                 nrow = 4, ncol = 3)

percent_conserv <- sum(survey[,1]) / sum(survey)
percent_cit_op <- sum(survey[1,]) / sum(survey)
percent_conserv_cit_op <- sum(survey[1,1]) / sum(survey)

percent_are_con_cit_op <- sum(survey[1,1]) / sum(survey[,1])
percent_are_mod_cit_op <- sum(survey[1,2]) / sum(survey[,2])
percent_are_lib_cit_op <- sum(survey[1,3]) / sum(survey[,3])
```

a. Percent conservative = 41%

b. Percent in favor of citizenship option = 31%

c. Percent conservative and in favor of citizenship option = 6%

d. Conservatives in favor of citizenship option = 15%

Moderates in favor of citizenship option = 33%

Liberals in favor of citizenship option = 58%

e. Political leaning and views on immigration do not appear to be independent. Conservatives, moderates, and liberals have very different amounts of support for the immigration law and there is a large sample size. We can get information on their views of immigration through their political leaning.

9.

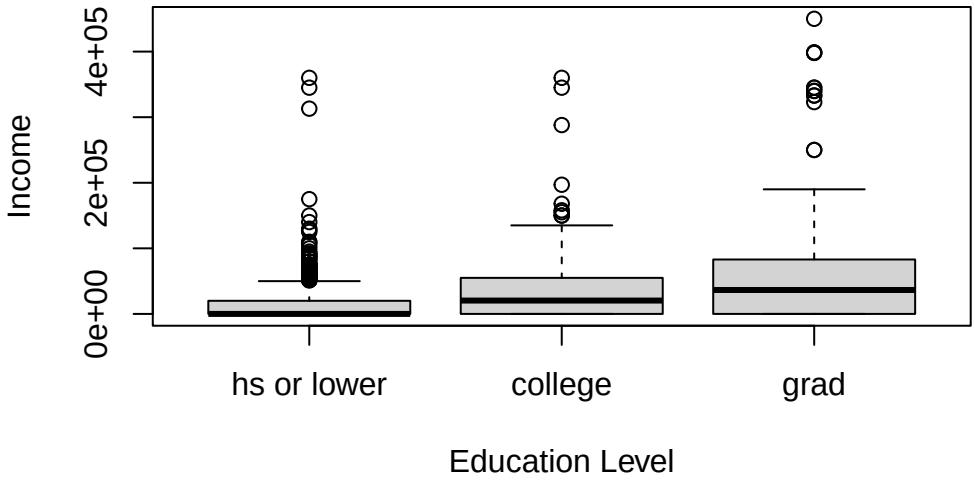
```
source("https://www.openintro.org/data/R/acs12.R")

median = median(acs12$income, na.rm = T)
```

```r
IQR = IQR(acs12$income, na.rm = T)

grad_prop <- table(acs12$edu)["grad"] / length(acs12$edu)

boxplot(acs12$income ~ acs12$edu, xlab = "Education Level", ylab = "Income")
```



Median = 3000

IQR = 3.37×10^4

Grad Proportion = 0.072